作者身份识别中不规范文本特征选择方法的研究*

郭　旭　祁瑞华
(大连外国语大学软件学院　大连 116044)

摘要: 【目的】从不规范文本中提取特征, 识别网络文本作者身份。【方法】提出两种在不规范文本中提取特征的方法: 利用在 Jaccard 系数的基础上定义的不规范文本相似度 M; 利用不规范文本在文本中出现的次数。【结果】两种特征的识别正确率分别达到 85.1%和 80.2%, 加入这两种特征后, 传统的基于统计值特征的分类器识别正确率分别提高 5.8%和 4%。【局限】只考虑到网络文本在词汇层面的不规范性, 并没有针对更高层面的特性进行研究, 如句法层面、结构层面。【结论】本文提出的特征提取方法, 可以有效地提取不规范文本特征, 有助于作者身份识别系统识别正确率的提升。

关键词: 作者身份　不规范文本　网络文本　文本相似度
分类号: TP391.1　G353

1　引　言

作者身份识别作为自然语言处理的一个重要方向, 一直倍受关注。随着微信、微博等社交网络的兴起与大数据时代的到来, 出于对道德与信息安全方面的考虑, 人们对网络文本作者身份正确认证的需求变得更加迫切。如在舆情监控中, 一些恶意信息是否出于同一作者, 又如垃圾邮件的作者归属问题。而在各种作者身份识别技术的应用中, 有一个问题一直对识别效果产生负面影响, 即“文本的书写不规范”。

传统的作者身份识别系统, 其识别正确率大都建立在样本书写规范的基础上。而当识别那些书写不规范的样本时, 各系统的识别正确率就会大幅下降。即使某些系统考虑了“书写不规范”的问题, 也只是简单的排除或还原, 并没有进行更深入的处理[1-2]。事实上, 不规范的文本表达往往是作者独特写作风格的集中体现。因此, 本文尝试从不规范文本中提取特征, 用于作者身份识别。这一提取特征的角度与网络应用环境紧密相联, 与现有的一些典型特征(如统计值特征、多层次特征)可以互补。因为, 不规范文本特征正是针对由于书写不规范而识别率较低的样本而提出的。

2　不规范文本的相关研究

不规范文本指由于作者的不规范或错误的书写行为而产生的文本。这些文本产生的原因主要有拼写错误、口语词汇、语气填补词、网络用语、表情符号、缩略语和俚语等[3]。在英文的书写环境下, 例如“lol”、“soooo”、“1ist”和“CU”等都属于不规范文本。不规范文本在正规出版物中很少出现, 但在以博客为代表的网络文本中却随处可见。因此, 在大多数针对网络文本开展的研究中, 都对不规范文本有所涉及, 主要研究内容可归纳为以下 4 个方面:

(1) 认为存在不规范文本是网络文本的特点, 但未针对不规范文本开展研究。目前多数研究虽然认为网络文本的不规范化对研究结果产生了负面影响, 但由于所设计的系统鲁棒性很强, 所以仍能取得很好的

通讯作者: 郭旭, ORCID: 0000-0002-0771-4845, E-mail: guoxu@dlufl.edu.cn。

*本文系国家社会科学基金项目“典籍英译国外读者网上评论观点挖掘研究”(项目编号: 15BYY028)和大连外国语大学科研项目“英文作者身份识别中书写不规范文本处理方法的研究”(项目编号: 2014XJQN15)的研究成果之一。

实验效果。因此, 并未针对不规范文本进行特定研究。如 Nie 等在进行超文本问答系统方面的研究时, 就提到了文本的拼写错误、俚语和缩写可能是影响分类正确率的原因[4]。

(2) 针对不规范文本进行实验, 取得了良好的实验结果。如陈叶旺等认为由于网络文本的不规范性使得其难以挖掘, 为此提出一种基于百度百科的网络文本语义主题抽取方法, 针对不规范文本进行实验。因为百度百科的词条内容丰富, 即使一些不规范的网络用语也包含其中, 所以无论是在规范的还是不规范的文本上进行实验, 都取得了很好的效果[5]。

(3) 将不规范文本还原为规范文本。如张文文等抽取文本观点句的实验中, 利用"校对词典"、"领域名词词典"和"网络情感词词典"将不规范文本还原为规范文本, 并标注其情感倾向[6]。又如 Dehkharghani 等在对 Twitter 上的文本进行情感分类时, 将特殊符号和表情符号还原为对应的情感[7]。

(4) 从不规范文本中提取特征。前三个方面, 虽然在一定程度上消除了不规范文本对相关研究所带来的负面影响, 但都没有充分利用不规范文本。实际上, 在以作者身份识别为代表的某些研究领域, 不规范文本往往可以提供很好的区分度。例如有些作者习惯使用网络用语 CU 表示 See You, 而另一些作者习惯在 So 后加多个 o 表示更强烈的情绪。因此, 如果可以找到从不规范文本中提取有效特征的方法, 用于作者身份识别。不仅可以消除不规范文本对识别带来的负面影响, 还可以有效地提高识别正确率。正如 Iqbal 等使用词汇拼写错误和句法错误特征识别电子邮件作者身份[8]。

3 不规范文本的相关定义

将一个单词表中的单词定义为规范文本。采用查表的方法, 找出文本中不包括在单词表中的, 且不属于命名实体、数字、网址等文本的单词, 以此作为不规范文本。同时定义了不规范度 N 和不规范文本相似度 M。

不规范度 N 表示文本的不规范程度, 公式如下:

$$N = S_n/S \tag{1}$$

其中, S_n 为样本中不规范文本数, S 为样本中单词的总数。

不规范文本相似度 M 表示文本之间的相似程度。传统的文本相似度计算方法大都是先通过某种算法(如 TF-IDF)获得文本的关键词, 从而将一段文本转换为由关键词组成的向量, 再利用余弦相似度或 Jaccard 系数表征相似度[9]。而本文在计算相似度 M 时, 以文本中包含的不规范文本作为关键词, 通过一种改进的 Jaccard 系数表征相似度。Jaccard 系数等于两个集合的交集除以两个集合的并集。本文的相似度计算也遵循这一思想, 但考虑到本文对相似度的应用并不是单纯地判断两个文本是否相似, 而是将同一文本与多个文本的相似度进行比较, 判断该文本与哪个文本更相似。因此, 在进行这种横向比较时, 应该满足如下三个条件:

(1) 随着两个文本之间"相同不规范文本"数量的增加, M 值增大。

(2) 随着某一"相同不规范文本"数量增加, M 值增大。但又不能过大, 避免某一"相同不规范文本"数量过大, 而使 M 值过大。因为, 在比较诸如文本间有两个出现一次的"相同不规范文本"和有一个出现两次的"相同不规范文本"时, 期望前者的 M 值大于后者。

(3) 避免某一文本中不规范文本过多或过少而带来的优势。

据此, 本文将文本 a 与文本 b 的不规范文本相似度 M 定义为:

$$M_{ab} = \frac{\sum_{i=1}^{n} \ln(P_{ai}+1) \times \ln(P_{bi}+1)}{\ln(S_a + S_b + 1)} \tag{2}$$

其中, n 为两个文本间不同的"相同不规范文本"的种类数, P_{ai} 为文本 a 中第 i 类"相同不规范文本"数, P_{bi} 为文本 b 中第 i 类"相同不规范文本"数, S_a 为文本 a 中不规范文本总数, S_b 为文本 b 中不规范文本总数。

公式(2)中, 计算所有"相同不规范文本"之和可以满足条件(1)。将某类"相同不规范文本"在各文本中的数目加 1 取自然对数可以满足条件(2), 这样在不规范文本总数相同的情况下, 一个出现多次的"相同不规范文本"和多个出现一次的"相同不规范文本"的对应关系为 $2^n–1$, 即一个出现 $2^n–1$ 次的"相同不规范文本"与 n 个出现一次的"相同不规范文本"所得 M 值相同。最后, 除以两个样本所包含的不规范文本数可以满足条件(3)。

4 不规范文本的获取

针对英文博客提取不规范文本, 数据主要来源于

两个语料库。其一是由 Schler 等构建的作者身份语料库[10-11]，该语料库包含 Blogger (https://blogger.com/)上的 19 320 位作者发布的 681 288 篇博文。其二是由 Ward 构建的包含大约 35 万单词的 Moby 单词表[12]。

在对 Moby 单词表进行扩展后，最终定义一个包含 377 121 个单词的单词表为规范文本。在此基础上，对作者身份语料库中 18 828 位作者发布的 517 643 篇博文提取不规范文本。具体流程如下:

(1) 文本预处理。使用斯坦福大学 NLP 小组开发的自然语言处理软件包[13]对语料库中的语料进行分词和词形还原的处理。

(2) 去除命名实体。使用斯坦福大学 NLP 小组开发的自然语言处理软件包[13]获得语料库中的命名实体，并将其去除。

(3) 初步获取不规范文本。采用查表法统计语料库中未出现在单词表中的单词与该单词出现的次数，以此作为初步的不规范文本。

(4) 筛选不规范文本。将初步不规范文本中的停用词删掉。

停用词表如表 1 所示:

表 1 停用词表

名称	示例
标点符号	单独出现的标点符号(如,.?); 连续多个标点符号(如???、***、^_^)不属于停用词。
撇号	i'm、n't
数字	12m、123
URL	http://blogger.com
邮箱地址	123@163.com
连字符	T-shirt
文件名	123.jpg
连接词	website、homepage
非英文	中文、韩文等非英语文本

最终得到不规范文本 193 028 种，共 1 365 942 个。其中完全由字母组成的不规范文本 182 549 种，共计 1 043 210 个，出现次数超过 1 次的不规范文本 65 529 种，共 1 238 443 个，出现次数超过 100 次的不规范文本 1 121 种，共 809 685 个，99%的不规范文本出现次数都不超过 60 次。出现次数前 10 的不规范文本，与出现次数前 10 的完全由字母组成的不规范文本如表 2 所示。

表 2 出现次数前 10 的两类不规范文本

名次	不规范文本	出现次数	不规范文本(字母)	出现次数
1	!!	66 991	lol	46 286
2	!!!	57 664	cuz	20 609
3	lol	46 286	hmm	12 236
4	:)	26 929	tt	8 273
5	cuz	20 609	juz	6 891
6	??	15 121	ang	6 822
7	!!!!	14 910	wif	6 437
8	hmm	12 236	omg	6 383
9	?!	10 253	sooo	6 258
10	???	9 729	liao	5 026

而在不规范度方面，所有文本的不规范度为 0.0109。按作者划分，每位作者的平均不规范度为 0.01282，标准差为 0.017，最高不规范度为 0.2268，最低不规范度为 0。其中，不规范度等于 0 的作者(即书写完全规范的作者)有 521 位，占作者总数的 2.78%。按博客划分，每篇博客的平均不规范度为 0.0155，标准差为 0.041，最高不规范度为 1，最低不规范度为 0。其中，不规范度大于 0 的博客(即存在不规范文本的博客)有 293 254 篇，占博客总数的 56.65%。

由此可见，在作者身份语料库中存在大量的不规范文本，可以用来进行进一步实验。同时，也在一定程度上证明了网络文本中普遍存在文本不规范的情况。

5 实验结果及分析

5.1 基于基本统计值特征的分类实验

为了判断不规范文本特征的有效性，与不规范文本特征进行对比实验，采用 18 种基本统计值作为特征识别作者身份，包括: 字符数、数字字符数、小写字母数、大写字母数、单词数、不同单词数、标点符号数、不同标点符号数、字长大于 4 的单词数、平均字长、出现一次单词数、出现二次单词数、出现二次以上单词数、句子数、平均句长(单词数)、平均句长(字符数)、最长句子单词数、最短句子单词数。

针对基本统计值特征的实验在数据挖掘软件 WEKA[1]下完成，分别采用贝叶斯网络(Bayes Network)、支持向量机(Support Vector Machine)、神经网络(Neural Networks)和套袋(Bagging)这 4 种分类器识别样本。其中，贝叶斯网络分类器基于树扩展朴素

贝叶斯 (Tree Augmented Naive Bayes, TAN)算法设计完成；支持向量机分类器基于 SMO(Sequential Minimal Optimization)算法设计完成，核函数选择 Polynomial Kernel; 神经网络分类器采用反向传播训练样本，学习率(Learning Rate)为 0.3，动量(Momentum)为 0.2，最高迭代次数为 500 次；套袋分类器以决策树作为基本模型，迭代 10 次后获得分类结果[14]。

在作者身份语料库中，随机抽取 10 组每组 6 位作者的实验样本，进行 10 折交叉验证实验。实验结果的平均值如表 3 所示:

表 3　实验结果平均值对照表

分类器	识别正确率(%)	用时(s)
贝叶斯网络	64.3	2
支持向量机	58.7	8
神经网络	67.8	177
套袋	67.3	9

根据上述实验结果，综合考虑识别正确率和用时，最终选择套袋分类器作为后续实验的分类器。

5.2　基于不规范文本相似度 M 的分类实验

该分类实验以不规范文本相似度 M 作为特征。需要说明的是，本实验在计算不规范文本相似度 M 时，考虑到诸如“sooooo”和“soooo”、“!!!”和“!!!!”属于同类不规范文本，因此采用类似词干还原的方法将其合并计算，即在计算不规范文本相似度 M 之前，“sooooo”和“soooo”都转换为“sooo”。

实验分别采用基于 K 近邻的贝叶斯算法和合一算法对样本进行分类。所谓基于 K 近邻的贝叶斯算法是指：分别计算未知样本与每一个已知样本的不规范文本相似度 M，并找到 M 值最大的 K 个样本，以每一类样本在 K 个样本中的比值作为类条件密度，最后利用贝叶斯公式判断未知样本的归属类别。所谓合一算法是指：把每一类中的所有已知样本合而为一个大的已知样本，计算未知样本与其的相似度 M，并将未知样本归属于 M 值最大的类别。

由于不规范文本相似度 M 表示文本间不规范文本方面的相似程度，所以对于那些没有出现不规范文本的样本(即不规范度 N 等于 0 的样本)，本文所设计的分类器并没有分类效果。此外，对于那些与所有已知样本没有相同的不规范文本的未知样本(即所有相似度 M 都等于 0 的样本)，本文所设计的分类器也没有很好的分类效果。因此，实验只对上述两类样本以外的样本有效，而样本有效率为有效样本数与所有样本数的比值。

根据上述算法分别设计相应的分类器，并从作者身份语料库中，随机选出文本数属于多、中、少的 6 位作者共计 1 652 个样本，进行 10 折交叉验证实验。具体实验数据如表 4 和表 5 所示。其中，表 4 记录了样本基本信息；表 5 是分别采用两种分类器识别有效文本的实验结果，该实验验证了在 6 位作者中选取 2 位作者的 15 种组合、选取 4 位作者的 15 种组合和选取 6 位作者的 1 种组合的分类效果。

表 4　样本信息表

作者编号	文本大小	样本数	作者编号	文本大小	样本数
A 1	434KB	844	A 4	143KB	73
A 2	409KB	362	A 5	75KB	178
A 3	157KB	127	A 6	36KB	68

表 5　实验结果平均值对照表

作者数	特征集	规范文本	无匹配文本	有效文本	有效率(%)	分类器	平均正确率(%)
2	不规范文本	29.3	7.2	18.5	33.6	K 近邻	95.0
						合一	95.4
	基本统计值	—	—	—	100	Bagging	88.1
4	不规范文本	58.5	13.7	37.9	34.3	K 近邻	88.5
						合一	89.2
	基本统计值	—	—	—	100	Bagging	75.4
6	不规范文本	87.8	19.5	57.9	35.0	K 近邻	83.2
						合一	85.1
	基本统计值	—	—	—	100	Bagging	66.8

为了更准确地验证不规范文本相似度 M 的识别效果，在作者身份语料库中，分别随机抽取 10 组每组 6 位作者、4 位作者和 2 位作者共 30 组数据，进行 10 折交叉验证，实验结果如表 6 所示。

通过上述对比实验可以发现，基于相似度 M 的分类器样本有效率在 40%左右。即，它对于大约 60%的样本是没有识别能力的，而这些样本中大部分属于文本规范的情况，另一小部分属于无相同不规范文本的情况。和大多数分类算法类似，两种分类器在 2 类问题上的识别效果最好，随着类别的增加识别正确率有所下降。此外，由于合一算法在进行决策时用

表 6 随机抽取样本实验结果平均值对照表

作者数	特征集	规范文本	无匹配文本	有效文本	有效率(%)	分类器	平均正确率(%)
2	不规范文本	4.5	1.8	4.8	43.3	K 近邻 合一	93.2 95.1
	基本统计值	—	—	—	100	Bagging	88.5
4	不规范文本	18.4	5.0	15.9	40.4	K 近邻 合一	84.4 86.7
	基本统计值	—	—	—	100	Bagging	74.6
6	不规范文本	23.3	6.6	23.8	44.3	K 近邻 合一	80.4 83.1
	基本统计值	—	—	—	100	Bagging	67.3

到的样本信息较多，所以除了个别样本外合一算法的识别效果要优于基于 K-近邻的贝叶斯算法，但后一种算法计算速度较快并具有更好的可扩展性，当有新的样本出现时，不再需要重复计算旧样本之间的相似度。

在对有效样本的识别中，基于不规范文本相似度 M 的分类器识别正确率平均要高于后者约 13%。随着类别数的增加，后者的识别正确率下降趋势较前者更迅速。这就使得，类别数越多以不规范文本相似度 M 为特征的分类器优势越明显。虽然样本有效率不高是不规范文本特征的先天缺点，但是能够明确地判断出可识别样本与不可识别样本这一特点，使其可以很容易与其他分类器相结合。

5.3 基于不规范文本出现次数的分类实验

除了以不规范文本相似度 M 为特征外，本文还以不规范文本在样本中出现的次数为特征。这样就可以把样本表示为一组由不规范文本出现次数组成的向量 $(N_1,N_2,N_3,\cdots,N_n)$。而特征向量的维度 n 由已知样本中不规范文本的种类决定。

首先针对 5.2 节中获取的 6 位作者进行分类实验。实验在数据挖掘软件 WEKA 下完成，采用以决策树作为基本模型的套袋分类器，并对比以基本统计值为特征的分类效果，实验结果如表 7 所示。可以发现，以不规范文本出现次数为特征和以不规范文本相似度 M 为特征的分类器，在样本有效率方面基本一致。因为两种特征针对的不规范文本种类是一样的，细微的差别在于后者将诸如“sooooo”和“soooo”这样的不规范文本进行合并。在识别正确率方面，前者略低于后者，但前者的特征形式更为标准，可以应用于绝大多数分类算法。此外，使用相同分类器的前提下，以不规范文本出现次数为特征的分类器，其识别正确率要明显高于基本统计值特征。

表 7 实验结果平均值对照表

作者数	特征集	规范文本	无匹配文本	有效文本	有效率(%)	平均正确率(%)
2	不规范文本	29.3	7.5	18.2	33.0	91.6
	基本统计值	—	—	—	100	88.1
4	不规范文本	58.5	14.2	37.3	33.8	83.8
	基本统计值	—	—	—	100	75.4
6	不规范文本	87.8	20.2	57.2	34.6	80.2
	基本统计值	—	—	—	100	66.8

5.4 不规范文本特征与基本统计值特征相结合的分类实验

克服不规范文本特征样本有效率不足的方法有两种：其一，将不规范文本次数特征与其他特征连接到一起形成一个新的特征向量；其二，先用基于不规范文本特征的分类器对有效样本分类，再用其他分类器分类无效样本。根据这两种方法，本文将不规范文本特征与基本统计值特征结合到一起识别作者身份。

根据方法一，将不规范文本特征与基本统计值特征连接到一起形成一个新的特征向量。并用套袋分类器对 5.2 节中获取的 6 位作者进行 10 折交叉验证实验，实验结果如表 8 所示：

表 8 实验结果平均值对照表

作者数	特征集	有效率(%)	正确率(%)	提升(%)
2	不规范文本 F1	33.0	91.6	0.9
	基本统计值 F2	100	88.1	
	F1+F2		89.0	
4	不规范文本 F1	33.8	83.8	1.6
	基本统计值 F2	100	75.4	
	F1+F2		77.0	
6	不规范文本 F1	34.6	80.2	4
	基本统计值 F2	100	66.8	
	F1+F2		70.8	

根据方法二，首先使用基于合一算法的以不规范文本相似度 M 为特征的分类器分类样本。再用以基本统计值为特征的 Bagging 分类器分类前者不能识别的无效样本。对 5.2 节中获取的 6 位作者进行 10 折交叉验证实验，实验结果如表 9 所示：

表 9　实验结果平均值对照表

作者数	特征集	有效率(%)	正确率(%)	提升(%)
2	不规范文本 F1	33.6	95.4	1.4
	基本统计值 F2	100	88.1	
	F1+F2		89.5	
4	不规范文本 F1	34.3	89.2	3.5
	基本统计值 F2	100	75.4	
	F1+F2		78.9	
6	不规范文本 F1	35.0	85.1	5.8
	基本统计值 F2	100	66.8	
	F1+F2		72.6	

为了进一步验证不规范文本特征的识别效果，实验又针对 5.2 节中随机抽取的 30 组数据进行 10 折交叉验证，实验结果如表 10 所示：

表 10　实验结果平均值对照表

作者数	特征集	有效率(%)	正确率(%)	提升(%)
2	不规范文本 F1	31.7	95.1	2.1
	基本统计值 F2	100	88.5	
	F1+F2		90.6	
4	不规范文本 F1	40.4	86.7	3.1
	基本统计值 F2	100	74.6	
	F1+F2		77.7	
6	不规范文本 F1	45.6	83.1	5.1
	基本统计值 F2	100	67.3	
	F1+F2		72.4	

通过上述实验可以发现，在引入不规范文本特征后识别正确率明显提高，最高提升 5.8%。可见，不规范文本特征可以有效地识别作者身份。相比于方法一，方法二的识别效果较好较稳定。因为方法二更能体现不规范文本特征能够明确地判断出可识别样本与不可识别样本这一特点。

6　结　语

针对网络文本大多书写不规范的特点，本文提出从不规范文本中提取特征以识别作者身份的方法。为此，本文设计并完成以相似度 M 为特征和以不规范文本次数为特征的作者识别实验。实验结果表明，基于不规范文本的特征可以有效地识别作者身份。所设计的特征模型与网络应用环境紧密相联，探索性地提出从不规范文本中提取特征来识别作者身份，这与现今大多数特征模型提取特征的角度不同。因此，它可以很容易地和其他特征模型相结合或作为其他特征模型的有效补充手段。

实验虽然在一定程度上验证了不规范文本用于作者识别的有效性，但还存在一些不足之处：

(1) 本文只是研究了作者身份识别中"属于哪个作者"这个问题，而没有涉及到"是否属于这个作者"的问题。这两个问题不同之处在于：前者可以是 2 类问题也可以是多类问题，在已知样本里各类别的样本都存在，并且未知样本必属于某一类别，其往往进行的是横向对比，判断未知样本更接近于哪一个已知样本；后者一般是判断"是"与"否"的两类问题，在已知样本里只有或只需要"是"的样本，通常需要给定一个阈值来判断未知样本是否属于已知样本。

(2) 只考虑到了网络文本在词汇层面的不规范性，并没有针对更高层面的特性进行研究，如句法层面、结构层面[15]。实际上，诸如"你走先"、"表告诉我"等在更高层面上的不规范书写行为，往往更能体现作者的写作习惯。

参考文献：

[1]　Abbasi A, Chen H. Applying Authorship Analysis to Extremist-group Web Forum Messages [J]. IEEE Intelligent Systems, 2005, 20(5): 67-75.

[2]　Iqbal F, Binsalleeh H, Fung B C M, et al. A Unified Data Mining Solution for Authorship Analysis in Anonymous Textual Communications [J]. Information Sciences, 2013, 231(9): 98-112.

[3]　骆昌日，何婷婷. 网络语言的特点及其情感性意义[J]. 武汉理工大学学报：社会科学版，2015, 28(2): 322-328. (Luo Changri, He Tingting. Characteristics of Internet Language and Its Emotional Meanings [J]. Journal of Wuhan University of Technology: Social Sciences Edition, 2015, 28(2): 322-328.)

[4]　Nie L, Wang M, Gao Y, et al. Beyond Text QA: Multimedia Answer Generation by Harvesting Web Information [J]. IEEE Transactions on Multimedia, 2013, 15(2): 426-441.

[5]　陈叶旺，王华珍，李海波,等. 基于百度百科与文本分类的

网络文本语义主题抽取方法[J]. 小型微型计算机系统, 2012, 33(12): 2605-2610. (Chen Yewang, Wang Huazhen, Li Haibo, et al. Topic Extraction Method for Chinese Web Text Based on Baidu Baike and Text Classification [J]. Journal of Chinese Computer Systems, 2012, 33(12): 2605-2610.)

[6] 张文文, 王挺. 不规范文本的无监督观点句抽取[J]. 计算机与数字工程, 2013, 41(1): 64-68. (Zhang Wenwen, Wang Ting. Unsupervised Subjective Sentence Extraction for Non-Standard Texts [J]. Computer and Digital Engineering, 2013, 41(1): 64-68.)

[7] Dehkharghani R, Mercan H, Javeed A, et al. Sentimental Causal Rule Discovery from Twitter [J]. Expert Systems with Applications, 2014, 41(10): 4950-4958.

[8] Iqbal F, Binsalleeh H, Fung B C M, et al. Mining Writeprints from Anonymous E-mails for Forensic Investigation [J]. Digital Investigation, 2010, 7(1): 56-64.

[9] 黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报, 2011, 34(5): 856-864. (Huang Chenghui, Yin Jian, Hou Fang. A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method [J]. Chinese Journal of Computers, 2011, 34(5): 856-864.)

[10] Schler J, Koppel M, Argamon S, et al. Effects of Age and Gender on Blogging [C]. In: Proceedings of the 2006 AAAI Spring Symposium. 2006.

[11] Schler J, Koppel M, Argamon S, et al. The Blog Authorship Corpus [DS/OL]. [2014-05-28]. http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm.

[12] Ward G. Moby Words [DS/OL]. [2016-06-24]. http://icon.shef.ac.uk/Moby/mwords.html.

[13] Manning C D, Surdeanu M, Bauer J, et al. The Stanford CoreNLP Natural Language Processing Toolkit [C]. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014.

[14] Witten I H, Frank E, Hall M A. Data Mining [M]. Beijing: China Machine Press, 2012.

[15] 祁瑞华, 杨德礼, 郭旭,等. 基于多层面文体特征的博客作者身份识别研究[J]. 情报学报, 2015,34(6):628-634. (Qi Ruihua, Yang Deli, Guo Xu, et al. Blogger Identification Based on Multidimensional Stylistic Features [J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(6): 628-634.)

作者贡献声明:

郭旭: 提出研究方案, 完成实验, 撰写论文;

祁瑞华: 提出研究思路和实验过程, 修改论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: guoxu@dlufl.edu.cn。

[1] 郭旭, 祁瑞华. nsw.xlxs. 不规范文本统计表.

[2] 郭旭, 祁瑞华. nsm.xlxs. 不规范文本识别效果统计表.

收稿日期: 2016-07-12
收修改稿日期: 2016-09-19

Using Non-standard Text Features to Identify Authors

Guo Xu　Qi Ruihua
(School of Software, Dalian University of Foreign Languages, Dalian 116044,China)

Abstract: [**Objective**] This paper aims to identify authors with features extracted from non-standard online texts. [**Methods**] First, we used the non-standard text similarity M defined by the Jaccard coefficient. Second, we adopted the frequency of non-standard text from the corpus. [**Results**] The recognition accuracy of the two features were 85.1% and 80.2%. Adding the two features to the traditional recognition mechanism, the precision of the system increased by 5.8% and 4%, respectively. [**Limitations**] We did not study the online texts from the syntactic and structure levels. [**Conclusions**] The proposed method could effectively extract the non-standard text features and then improve the accuracy of author identification.

Keywords: Author identification　Non-standard text　Network text　Text similarity